AN ENHANCED TRAINING-FREE ASD METHOD FOR NOISE-AWARE UNSUPERVISED ANOMALOUS SOUND DETECTION FOR MACHINE CONDITION MONITORING

Technical Report

Qin Xie, Luowei Ma, Shiyang Pei, and *Qinghua Huang*

ShangHai University
(2156381275@qq.com; shu_mlw@163.com;
2815429655@qq.com; qinghua@shu.edu.cn)

ABSTRACT

This report describes our method for DCASE 2026 Task 2: Noise-aware Unsupervised Anomalous Sound Detection (ASD) for Machine Condition Monitoring. Our method follows a training-free framework based on frozen audio embeddings and memory-bank-based anomaly scoring. To better utilize the noise-aware dual-channel recordings provided in the challenge, we introduce a difference-based feature enhancement strategy that combines near-field and far-field audio information. Furthermore, we investigate several temporal pooling methods to improve the aggregation of frame-level embeddings extracted by a frozen Efficient Audio Transformer (EAT). Experimental results show that both the proposed feature enhancement and temporal pooling strategies provide consistent improvements over the baseline system. The final system achieves enhanced anomaly detection performance while preserving the simplicity of training-free ASD methods.

Index Terms—anomalous sound detection, noise-aware learning, dual-channel audio, training-free learning, temporal pooling

1. INTRODUCTION

Anomalous sound detection aims to identify abnormal operating conditions of industrial machines using only reference recordings collected under normal conditions. As a representative benchmark in this field, DCASE Task 2 has continuously promoted the development of ASD systems toward increasingly realistic and challenging scenarios [1].

Compared with previous editions, DCASE 2026 Task 2 introduces synchronized near-field and far-field recordings. The near-field channel is expected to capture machine-related acoustic events more clearly, whereas the far-field channel may contain additional environmental information. Effectively exploiting the complementary information provided by these two recording conditions therefore becomes an important challenge for ASD systems.

GenRep has demonstrated that generic audio representations extracted from frozen audio encoders can achieve competitive ASD performance without task-specific model adaptation [2]. However, the original GenRep framework was designed primarily for single-channel audio and typically employs mean pooling for clip-level embedding aggregation. To address these limitations, we introduce two modifications to the GenRep framework. First, we propose a dual-channel difference representation to better exploit the information contained in synchronized near-field and far-field recordings. Second, we investigate multiple temporal pooling strategies for aggregating frame-level embeddings into clip-level representations.

Experimental results show that the proposed dual-channel representation and temporal pooling strategy consistently improve anomaly detection performance.

2. METHOD

2.1. Overall Framework

The proposed method is built upon the GenRep framework and introduces two key modifications: dual-channel feature enhancement and temporal pooling.

Given a pair of synchronized near-field and far-field recordings provided by DCASE2026 Task 2, we first construct an enhanced audio representation using a difference-based dual-channel fusion strategy. The enhanced signal is then fed into a frozen EAT encoder to extract frame-level audio embeddings. Subsequently, temporal pooling is applied to aggregate the frame-level representations into a clip-level embedding. Finally, anomaly scores are computed using source-domain and target-domain memory banks together with density-based score normalization.

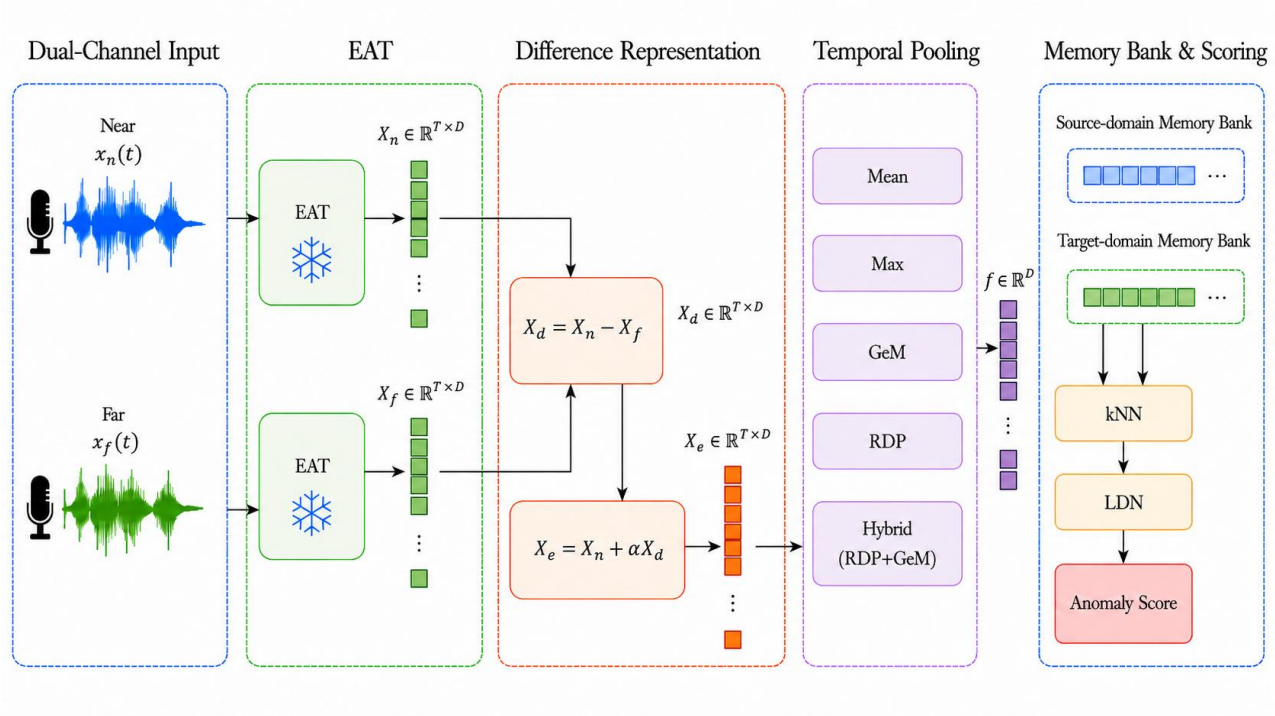

Figure 1: Overview of the proposed noise-aware training-free anomalous sound detection framework.

Compared with the original GenRep framework, our method introduces two modifications:

- A dual-channel difference representation is employed to better exploit the information provided by near-field and far-field microphones.
- Multiple temporal pooling strategies are investigated to improve the aggregation of frame-level embeddings.

2.2. Dual-Channel Difference Representation

Unlike previous DCASE ASD datasets that provide only a single audio channel, DCASE2026 introduces synchronized near-field and far-field recordings. The near-field microphone captures machine-related acoustic characteristics more clearly, while the far-field microphone contains additional environmental noise and propagation information.

To better exploit the complementary information from both channels, we construct a difference signal between the near-field and far-field recordings:

$$X_d = X_n - X_f \tag{1}$$

where $X_n$and X_f denote the near-field and far-field audio signals, respectively.
The final input signal is defined as

$$X_e = X_n + \alpha X_d \tag{2}$$

The proposed difference-based representation is designed to better exploit the complementary information contained in the near-field and far-field recordings. Since machine-related acoustic events are generally more prominent in the near-field channel, while environmental sounds are shared across both channels, the difference signal provides additional cues about machine-specific characteristics. By combining the near-field signal with the channel difference, the proposed representation enhances discriminative acoustic information for anomaly detection.

2.3. Audio Embedding Extraction

For feature extraction, we employ the Efficient Audio Transformer [3] as the backbone encoder. The EAT model is kept completely frozen during both training and inference stages. Given an input audio clip, EAT outputs a sequence of frame-level embeddings, where each embedding corresponds to a specific temporal segment of the input. Following the training-free philosophy of GenRep, no parameter updating or fine-tuning is performed on the pre-trained model, and all anomaly detection decisions are made solely based on the extracted embeddings.

2.4. Temporal Pooling Strategies

The frozen EAT encoder produces a sequence of frame-level embeddings for each audio recording. Before anomaly scoring can be performed, these frame-level representations must be aggregated into a fixed-dimensional clip-level embedding.

Previous studies have demonstrated that the selection of temporal pooling strategies can substantially affect the quality of audio representations and, consequently, the performance of training-free anomalous sound detection systems [4]. Therefore, several pooling methods were investigated within the proposed framework to assess their effectiveness in aggregating frame-level embeddings.

Based on the experimental results, Hybrid Pooling was selected as the final temporal aggregation strategy. Through the combination of adaptive frame weighting and feature aggregation, informative acoustic events are emphasized while anomaly-related

characteristics are better preserved. As a result, more discriminative clip-level representations can be obtained for subsequent anomaly detection.

The aggregated embeddings are then stored in the memory banks and are used for nearest-neighbor-based anomaly score computation. The effectiveness of the selected pooling strategy is further analyzed in the experimental section.

3. EXPERIMENTAL RESULTS

The experimental results are summarized in Table 1. The proposed method achieves an overall official score of 0.6254. Among all machine types, ValveEmu achieves the best performance, obtaining an official score of 0.8643, while Fan remains the most challenging category. Overall, the results demonstrate that the proposed dual-channel feature enhancement and temporal pooling strategies can effectively improve anomaly detection performance under the noise-aware setting.

Table 1: Performance of the proposed method on each machine type.

Machine Type	Source AUC	Source AUC	pAUC	Official Score
ToyCar	0.6956	0.7710	0.5684	0.7333
ToyCarEmu	0.6164	0.7288	0.5753	0.6726
BearingEmu	0.5590	0.6156	0.5305	0.5873
Fan	0.5446	0.5742	0.5063	0.5594
GearboxEmu	0.6778	0.5934	0.5268	0.6356
SliderEmu	0.6460	0.7088	0.5211	0.6774
ValveEmu	0.8362	0.8924	0.8011	0.8643

4. DISCUSSION AND CONCLUSION

This report presented our method to DCASE 2026 Task 2. Built upon the training-free GenRep framework, the proposed system incorporates a dual-channel feature enhancement strategy and temporal pooling for audio representation aggregation.

Experimental results demonstrate the effectiveness of the proposed approach, achieving an official score of 0.6254 while maintaining the simplicity and robustness of training-free ASD methods.

Future work will investigate more effective dual-channel representations and feature aggregation strategies for industrial anomalous sound detection.

5. REFERENCES

[1] T. Nishida et al., “Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” arXiv preprint arXiv:2506.10097, 2025.

[2] P. Saengthong and T. Shinozaki, “GENREP for first-shot unsupervised anomalous sound detection of DCASE 2025 challenge,” DCASE 2025 Challenge Technical Report, 2025.

[3] C.W. Wu et al., “Efficient audio transformer: An efficient audio transformer architecture for self-supervised learning of audio representations,” in Proc. ICASSP, 2022.

[4] K. Wilkinghoff, S. Yadav, and Z.-H. Tan, “Temporal pooling strategies for training-free anomalous sound detection with self-supervised audio embeddings,” arXiv preprint arXiv:2603.04605, 2026.